Text as Data

Justin Grimmer

Associate Professor
Department of Political Science
Stanford University

October 28th, 2014

"Vanilla" Latent Dirichlet Allocation

1) Task:
   - Discover thematic content of documents
   - Quickly explore documents
2) Objective Function

$$f(\boldsymbol{X}, \boldsymbol{\pi}, \boldsymbol{\Theta}, \boldsymbol{\alpha})$$

Where:
   - $\boldsymbol{\pi} = N \times K$ matrix with row $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{iK}) \rightsquigarrow$ proportion of a document allocated to each topic
   - $\boldsymbol{\Theta} = K \times J$ matrix, with row $\boldsymbol{\theta}_k = (\theta_{1k}, \theta_{2k}, \ldots, \theta_{kJ}) \rightsquigarrow$topics
   - $\boldsymbol{\alpha} = K$ element long vector, population prior for $\boldsymbol{\pi}$.
3) Optimization
   - Variational Approximation$\rightsquigarrow$ EM Algorithm where every step is an "E"
   - Collapsed Gibbs Sampling$\rightsquigarrow$ MCMC algorithm
   - Many other variants
4) Validation$\rightsquigarrow$ many of the same methods from clustering

Topic and Mixed Membership Models

Clustering

Document $\rightsquigarrow$ One Cluster

Doc 1

Doc 2

Doc 3

$\vdots$

Doc N

Cluster 1

Cluster 2

$\vdots$

Cluster K

Topic and Mixed Membership Models

Clustering

Document $\rightsquigarrow$ One Cluster

Doc 1

Doc 2

Doc 3

$\vdots$

Doc N

Cluster 1

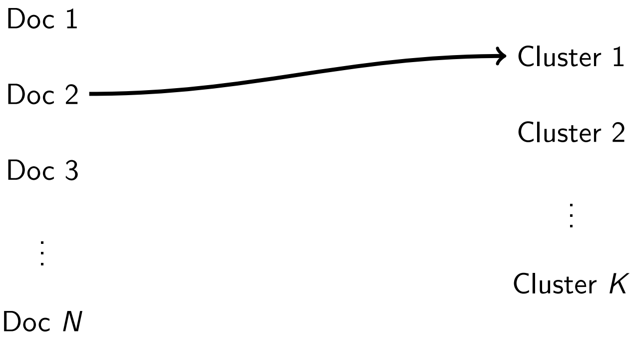
Cluster 2

$\vdots$

Cluster K

Topic and Mixed Membership Models

Clustering
Document $\rightsquigarrow$ One Cluster



Topic and Mixed Membership Models

Clustering

Document $\rightsquigarrow$ One Cluster

Doc 1

Doc 2

Doc 3

$\vdots$

Doc N

Cluster 1

Cluster 2

$\vdots$

Cluster K

Topic and Mixed Membership Models

Clustering

Document $\rightsquigarrow$ One Cluster

Doc 1 | Cluster 1
Doc 2 | Cluster 2
Doc 3 | $\vdots$
$\vdots$ | Cluster K
Doc N

Topic and Mixed Membership Models

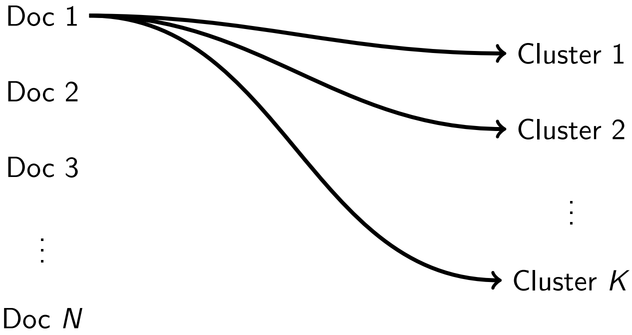
Topic Models (Mixed Membership)
Document $\rightsquigarrow$ Many clusters

Doc 1	
	Cluster 1
Doc 2	
	Cluster 2
Doc 3	
	$\vdots$
$\vdots$	
	Cluster K
Doc N	

Topic and Mixed Membership Models

Topic Models (Mixed Membership)

Document $\rightsquigarrow$ Many clusters

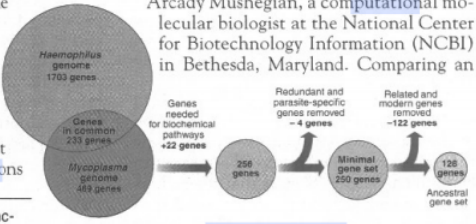


A Statistical Highlighter (With Many Colors)

Seeking Life's Bare (Genetic) Necessities

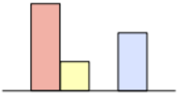
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.



Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

- Consider document i, $(i = 1, 2, \ldots, N)$.

Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

- Consider document i, $(i = 1, 2, \ldots, N)$.
- Suppose there are M_i total words and $\boldsymbol{x}_i$ is an $M_i \times 1$ vector, where x_{im} describes the m^{th} word used in the document*.

Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

- Consider document i, $(i = 1, 2, \ldots, N)$.
- Suppose there are M_i total words and $\boldsymbol{x}_i$ is an $M_i \times 1$ vector, where x_{im} describes the m^{th} word used in the document*.

*Notice: this is a different representation than a document-term matrix. x_{im} is a number that says which of the J words are used. The difference is for clarity and we'll this representation is closely related to document-term matrix

Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

- Consider document i, $(i = 1, 2, \ldots, N)$.
- Suppose there are M_i total words and $\boldsymbol{x}_i$ is an $M_i \times 1$ vector, where x_{im} describes the m^{th} word used in the document*.

$$\boldsymbol{\pi}_i | \boldsymbol{\alpha} \sim \text{Dirichlet}(\boldsymbol{\alpha})$$

Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

- Consider document i, $(i = 1, 2, \ldots, N)$.
- Suppose there are M_i total words and $\boldsymbol{x}_i$ is an $M_i \times 1$ vector, where x_{im} describes the m^{th} word used in the document*.

$$
\begin{aligned}
\boldsymbol{\pi}_i | \boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{\tau}_{im} | \boldsymbol{\pi}_i &\sim \text{Multinomial}(1, \boldsymbol{\pi}_i)
\end{aligned}
$$

Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

- Consider document i, $(i = 1, 2, \ldots, N)$.
- Suppose there are M_i total words and $\boldsymbol{x}_i$ is an $M_i \times 1$ vector, where x_{im} describes the m^{th} word used in the document*.

$$
\begin{aligned}
\boldsymbol{\pi}_i | \boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{\tau}_{im} | \boldsymbol{\pi}_i &\sim \text{Multinomial}(1, \boldsymbol{\pi}_i) \\
x_{im} | \boldsymbol{\theta}_k, \tau_{imk} = 1 &\sim \text{Multinomial}(1, \boldsymbol{\theta}_k)
\end{aligned}
$$

Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

- Consider document i, $(i = 1, 2, \ldots, N)$.
- Suppose there are M_i total words and $\boldsymbol{x}_i$ is an $M_i \times 1$ vector, where x_{im} describes the m^{th} word used in the document*.

$$\boldsymbol{\theta}_k \sim \text{Dirichlet}(\mathbf{1})$$

$$\begin{aligned}
\boldsymbol{\pi}_i|\boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{\tau}_{im}|\boldsymbol{\pi}_i &\sim \text{Multinomial}(1, \boldsymbol{\pi}_i) \\
x_{im}|\boldsymbol{\theta}_k, \tau_{imk} = 1 &\sim \text{Multinomial}(1, \boldsymbol{\theta}_k)
\end{aligned}$$

Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

- Consider document i, $(i = 1, 2, \ldots, N)$.
- Suppose there are M_i total words and $\boldsymbol{x}_i$ is an $M_i \times 1$ vector, where x_{im} describes the m^{th} word used in the document*.

$$
\begin{aligned}
\boldsymbol{\theta}_k &\sim \text{Dirichlet}(\mathbf{1}) \\
\alpha_k &\sim \text{Gamma}(\alpha, \beta) \\
\boldsymbol{\pi}_i | \boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{\tau}_{im} | \boldsymbol{\pi}_i &\sim \text{Multinomial}(1, \boldsymbol{\pi}_i) \\
x_{im} | \boldsymbol{\theta}_k, \tau_{imk} = 1 &\sim \text{Multinomial}(1, \boldsymbol{\theta}_k)
\end{aligned}
$$

Vanilla Latent Dirichlet Allocation$\rightsquigarrow$ Objective Function

Together the model implies the following posterior:

Vanilla Latent Dirichlet Allocation$\rightsquigarrow$ Objective Function

Together the model implies the following posterior:

$$p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \alpha|\boldsymbol{X}) \quad \propto \quad p(\boldsymbol{\alpha})p(\boldsymbol{\pi}|\alpha)p(\boldsymbol{T}|\boldsymbol{\pi})p(\boldsymbol{X}|\theta, \boldsymbol{T})$$

Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

Together the model implies the following posterior:

$$
\begin{aligned}
p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \alpha | \boldsymbol{X}) & \propto p(\boldsymbol{\alpha}) p(\boldsymbol{\pi} | \alpha) p(\boldsymbol{T} | \boldsymbol{\pi}) p(\boldsymbol{X} | \theta, \boldsymbol{T}) \\
& \propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[p(\boldsymbol{\pi}_i | \boldsymbol{\alpha}) \prod_{m=1}^{M_i} p(\boldsymbol{\tau}_{im} | \boldsymbol{\pi}) p(x_{im} | \boldsymbol{\theta}_k, \tau_{imk} = 1) \right]
\end{aligned}
$$

Vanilla Latent Dirichlet Allocation$\rightsquigarrow$ Objective Function

Together the model implies the following posterior:

$$\begin{aligned}
p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha}|\boldsymbol{X}) &\propto p(\boldsymbol{\alpha})p(\boldsymbol{\pi}|\boldsymbol{\alpha})p(\boldsymbol{T}|\boldsymbol{\pi})p(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{T}) \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[p(\boldsymbol{\pi}_i|\boldsymbol{\alpha}) \prod_{m=1}^{M_i} p(\boldsymbol{\tau}_{im}|\boldsymbol{\pi}) p(x_{im}|\boldsymbol{\theta}_k, \tau_{imk} = 1) \right] \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_{ik}^{\alpha_k - 1} \prod_{m=1}^{M} \prod_{k=1}^{K} \left[\pi_{ik} \prod_{j=1}^{J} \theta_{jk}^{x_{imj}} \right]^{\tau_{ikm}} \right]
\end{aligned}$$

Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

Together the model implies the following posterior:

$$
\begin{aligned}
p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha} | \boldsymbol{X}) &\propto p(\boldsymbol{\alpha}) p(\boldsymbol{\pi} | \boldsymbol{\alpha}) p(\boldsymbol{T} | \boldsymbol{\pi}) p(\boldsymbol{X} | \boldsymbol{\theta}, \boldsymbol{T}) \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[p(\boldsymbol{\pi}_i | \boldsymbol{\alpha}) \prod_{m=1}^{M_i} p(\boldsymbol{\tau}_{im} | \boldsymbol{\pi}) p(x_{im} | \boldsymbol{\theta}_k, \tau_{imk} = 1) \right] \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_{ik}^{\alpha_k - 1} \prod_{m=1}^{M} \prod_{k=1}^{K} \left[\pi_{ik} \prod_{j=1}^{J} \theta_{jk}^{x_{imj}} \right]^{\tau_{ikm}} \right]
\end{aligned}
$$

Vanilla Latent Dirichlet Allocation⇝ Objective Function

Together the model implies the following posterior:

$$\begin{aligned}
p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha}|\boldsymbol{X}) &\propto p(\boldsymbol{\alpha})p(\boldsymbol{\pi}|\boldsymbol{\alpha})p(\boldsymbol{T}|\boldsymbol{\pi})p(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{T}) \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[p(\boldsymbol{\pi}_i|\boldsymbol{\alpha}) \prod_{m=1}^{M_i} p(\boldsymbol{\tau}_{im}|\boldsymbol{\pi}) p(x_{im}|\boldsymbol{\theta}_k, \tau_{imk} = 1) \right] \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_{ik}^{\alpha_k - 1} \prod_{m=1}^{M} \prod_{k=1}^{K} \left[\pi_{ik} \prod_{j=1}^{J} \theta_{jk}^{x_{imj}} \right]^{\tau_{ikm}} \right]
\end{aligned}$$

Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

Together the model implies the following posterior:

$$
\begin{aligned}
p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha} | \boldsymbol{X}) & \propto p(\boldsymbol{\alpha}) p(\boldsymbol{\pi}|\boldsymbol{\alpha}) p(\boldsymbol{T}|\boldsymbol{\pi}) p(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{T}) \\
& \propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[p(\boldsymbol{\pi}_i|\boldsymbol{\alpha}) \prod_{m=1}^{M_i} p(\boldsymbol{\tau}_{im}|\boldsymbol{\pi}) p(x_{im}|\boldsymbol{\theta}_k, \tau_{imk} = 1) \right] \\
& \propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_{ik}^{\alpha_k - 1} \prod_{m=1}^{M} \prod_{k=1}^{K} \left[\pi_{ik} \prod_{j=1}^{J} \theta_{jk}^{x_{imj}} \right]^{\tau_{ikm}} \right]
\end{aligned}
$$

Optimization:

Vanilla Latent Dirichlet Allocation$\leadsto$ Objective Function

Together the model implies the following posterior:

$$
\begin{aligned}
p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha}|\boldsymbol{X}) &\propto p(\boldsymbol{\alpha})p(\boldsymbol{\pi}|\boldsymbol{\alpha})p(\boldsymbol{T}|\boldsymbol{\pi})p(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{T}) \\
&\propto p(\boldsymbol{\alpha})\prod_{i=1}^{N}\left[p(\boldsymbol{\pi}_i|\boldsymbol{\alpha})\prod_{m=1}^{M_i} p(\boldsymbol{\tau}_{im}|\boldsymbol{\pi})p(x_{im}|\boldsymbol{\theta}_k, \tau_{imk} = 1)\right] \\
&\propto p(\boldsymbol{\alpha})\prod_{i=1}^{N}\left[\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\pi_{ik}^{\alpha_k-1}\prod_{m=1}^{M}\prod_{k=1}^{K}\left[\pi_{ik}\prod_{j=1}^{J}\theta_{jk}^{x_{imj}}\right]^{\tau_{ikm}}\right]
\end{aligned}
$$

Optimization:

- Variational Approximation$\leadsto$ Find "closest" distribution

Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

Together the model implies the following posterior:

$$
\begin{aligned}
p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha} | \boldsymbol{X}) & \propto p(\boldsymbol{\alpha}) p(\boldsymbol{\pi} | \boldsymbol{\alpha}) p(\boldsymbol{T} | \boldsymbol{\pi}) p(\boldsymbol{X} | \boldsymbol{\theta}, \boldsymbol{T}) \\
& \propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[p(\boldsymbol{\pi}_i | \boldsymbol{\alpha}) \prod_{m=1}^{M_i} p(\boldsymbol{\tau}_{im} | \boldsymbol{\pi}) p(x_{im} | \boldsymbol{\theta}_k, \tau_{imk} = 1) \right] \\
& \propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_{ik}^{\alpha_k - 1} \prod_{m=1}^{M} \prod_{k=1}^{K} \left[\pi_{ik} \prod_{j=1}^{J} \theta_{jk}^{x_{imj}} \right]^{\tau_{ikm}} \right]
\end{aligned}
$$

Optimization:

- Variational Approximation $\rightsquigarrow$ Find "closest" distribution
- Gibbs sampling $\rightsquigarrow$ MCMC algorithm to approximate posterior

Vanilla Latent Dirichlet Allocation$\rightsquigarrow$ Objective Function

Together the model implies the following posterior:

$$
\begin{aligned}
p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha}|\boldsymbol{X}) &\propto p(\boldsymbol{\alpha})p(\boldsymbol{\pi}|\boldsymbol{\alpha})p(\boldsymbol{T}|\boldsymbol{\pi})p(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{T}) \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[p(\boldsymbol{\pi}_i|\boldsymbol{\alpha}) \prod_{m=1}^{M_i} p(\boldsymbol{\tau}_{im}|\boldsymbol{\pi}) p(x_{im}|\boldsymbol{\theta}_k, \tau_{imk} = 1) \right] \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_{ik}^{\alpha_k - 1} \prod_{m=1}^{M} \prod_{k=1}^{K} \left[\pi_{ik} \prod_{j=1}^{J} \theta_{jk}^{x_{imj}} \right]^{\tau_{ikm}} \right]
\end{aligned}
$$

Optimization:

- Variational Approximation$\rightsquigarrow$ Find "closest" distribution
- Gibbs sampling $\rightsquigarrow$ MCMC algorithm to approximate posterior

Described in the slides appendix

Running a Topic Model with Mallet

`to the Mallet/R Code!!`

Why does this work $\rightsquigarrow$ Co-occurrence

Where's the information for each word's topic?

Why does this work $\rightsquigarrow$ Co-occurrence

Where's the information for each word's topic?
Reconsider document-term matrix

Why does this work $\rightsquigarrow$ Co-occurrence

Where's the information for each word's topic?

Reconsider document-term matrix

	Word_1	Word_2	$\ldots$	Word_J
Doc_1	0	1	$\ldots$	0
Doc_2	2	0	$\ldots$	3
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Doc_N	0	1	$\ldots$	1

Why does this work $\rightsquigarrow$ Co-occurrence

Where's the information for each word's topic?
Reconsider document-term matrix

	Word_1	Word_2	$\ldots$	Word_J
Doc_1	0	1	$\ldots$	0
Doc_2	2	0	$\ldots$	3
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Doc_N	0	1	$\ldots$	1

Inner product of Documents (rows): $\mathbf{Doc}_i^{'}\mathbf{Doc}_l$

Why does this work $\rightsquigarrow$ Co-occurrence

Where's the information for each word's topic?

Reconsider document-term matrix

	Word_1	Word_2	$\ldots$	Word_J
Doc_1	0	1	$\ldots$	0
Doc_2	2	0	$\ldots$	3
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Doc_N	0	1	$\ldots$	1

Inner product of Documents (rows): $\textbf{Doc}_i^{'}\textbf{Doc}_l$

Inner product of Terms (columns): $\textbf{Word}_j^{'}\textbf{Word}_k$

Why does this work $\rightsquigarrow$ Co-occurrence

Where's the information for each word's topic?

Reconsider document-term matrix

	Word_1	Word_2	$\ldots$	Word_J
Doc_1	0	1	$\ldots$	0
Doc_2	2	0	$\ldots$	3
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Doc_N	0	1	$\ldots$	1

Inner product of Documents (rows): $\textbf{Doc}_i^{'}\textbf{Doc}_l$

Inner product of Terms (columns): $\textbf{Word}_j^{'}\textbf{Word}_k$

Allows: measure of correlation of term usage across documents

(heuristically: partition words, based on usage in documents)

Why does this work $\rightsquigarrow$ Co-occurrence

Where's the information for each word's topic?

Reconsider document-term matrix

	Word_1	Word_2	$\ldots$	Word_J
Doc_1	0	1	$\ldots$	0
Doc_2	2	0	$\ldots$	3
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Doc_N	0	1	$\ldots$	1

Inner product of Documents (rows): $\textbf{Doc}_i^{'}\textbf{Doc}_l$

Inner product of Terms (columns): $\textbf{Word}_j^{'}\textbf{Word}_k$

Allows: measure of correlation of term usage across documents (heuristically: partition words, based on usage in documents)

Latent Semantic Analysis: Reduce information in matrix using linear algebra (provides similar results, difficult to generalize)

Why does this work $\rightsquigarrow$ Co-occurrence

Where's the information for each word's topic?

Reconsider document-term matrix

	Word_1	Word_2	$\ldots$	Word_J
Doc_1	0	1	$\ldots$	0
Doc_2	2	0	$\ldots$	3
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Doc_N	0	1	$\ldots$	1

Inner product of Documents (rows): $\mathbf{Doc}_i^{'}\mathbf{Doc}_l$

Inner product of Terms (columns): $\mathbf{Word}_j^{'}\mathbf{Word}_k$

Allows: measure of correlation of term usage across documents (heuristically: partition words, based on usage in documents)

Latent Semantic Analysis: Reduce information in matrix using linear algebra (provides similar results, difficult to generalize)

Biclustering: Models that partition documents and words simultaneously

Why does this work $\rightsquigarrow$ Co-occurrence logic (h/t Colorado Reed Tutorial)

$$p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \alpha | \boldsymbol{X}) \quad \propto \quad p(\boldsymbol{\alpha}) p(\boldsymbol{\pi} | \boldsymbol{\alpha}) p(\boldsymbol{T} | \boldsymbol{\pi}) p(\boldsymbol{X} | \boldsymbol{\theta}, \boldsymbol{T})$$

Why does this work$\rightsquigarrow$ Co-occurrence logic (h/t Colorado Reed Tutorial)

$$p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha}|\boldsymbol{X}) \propto p(\boldsymbol{\alpha})p(\boldsymbol{\pi}|\boldsymbol{\alpha})p(\boldsymbol{T}|\boldsymbol{\pi})\underbrace{p(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{T})}_{1}$$

1) $\boldsymbol{\theta} \rightsquigarrow$ Greater weight on terms that occur together

Why does this work $\rightsquigarrow$ Co-occurrence logic (h/t Colorado Reed Tutorial)

$$p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \alpha | \boldsymbol{X}) \quad \propto \quad p(\boldsymbol{\alpha}) p(\boldsymbol{\pi} | \boldsymbol{\alpha}) \underbrace{p(\boldsymbol{T} | \pi)}_{2} \underbrace{p(\boldsymbol{X} | \theta, \boldsymbol{T})}_{1}$$

1) $\boldsymbol{\theta} \rightsquigarrow$ Greater weight on terms that occur together
2) $\boldsymbol{\pi} \rightsquigarrow$ Greater weight on indicators that appear more regularly

Why does this work $\rightsquigarrow$ Co-occurrence logic (h/t Colorado Reed Tutorial)

$$p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \alpha|\boldsymbol{X}) \propto p(\alpha)\underbrace{p(\boldsymbol{\pi}|\alpha)}_{3}\underbrace{p(\boldsymbol{T}|\pi)}_{2}\underbrace{p(\boldsymbol{X}|\theta, \boldsymbol{T})}_{1}$$

1) $\boldsymbol{\theta} \rightsquigarrow$ Greater weight on terms that occur together
2) $\boldsymbol{\pi} \rightsquigarrow$ Greater weight on indicators that appear more regularly
3) $\boldsymbol{\alpha} \rightsquigarrow$ Emphasis on $\boldsymbol{\pi}$ with greater weight

Validation$\rightsquigarrow$ Topic Intrusion

Thursday$\rightsquigarrow$ discussed several validations

- Labeling paragraphs
  - Identify separating words automatically
  - Label topics manually (read!)
- Statistical methods
  1) Entropy
  2) Exclusivity
  3) Cohesiveness
- Experiment Based Methods
  - Word intrusion$\rightsquigarrow$ topic validity
  - Topic intrusion$\rightsquigarrow$ model fit

Validation$\rightsquigarrow$ Topic Intrusion

1) Ask research assistant to read paragraph
2) Construct experiment
    - For the document, select top three topics
    - Select a fourth topic
    - Show participant, ask her/him to identify intruder

Higher identification$\rightsquigarrow$ topics are a better model of text

Example 1: Japanese Campaign Manifestos (Catalinac 2011)

- Why is Japan revising its constitution?
- IR question: why is Japan now willing to engage militaristic foreign action?
- One explanation: election reform in 1993, changed electoral incentives
- To answer well: characterize campaigns across 50 + years
  - That sounds hard
  - That sounds impossible
- Determined (relentless) data collection
- Latent Dirichlet Allocation (on japanese texts)

Example 1: Japanese Campaign Manifestos (Catalinac 2011)

Japanese Elections:

Example 1: Japanese Campaign Manifestos (Catalinac 2011)

Japanese Elections:

- Election Administration Commission runs elections $\rightarrow$ district level

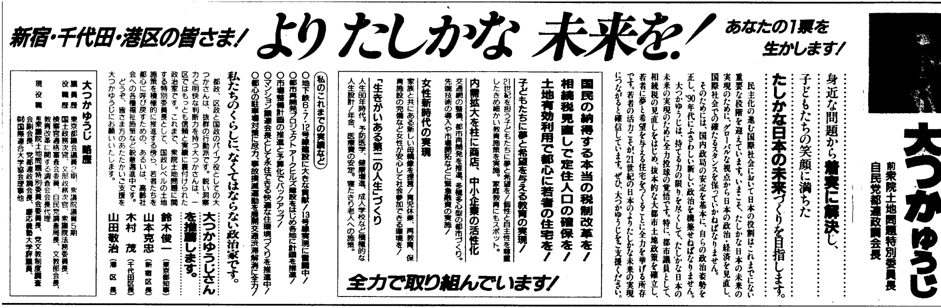
Example 1: Japanese Campaign Manifestos (Catalinac 2011)

Japanese Elections:

- Election Administration Commission runs elections $\rightarrow$ district level
- Required to submit manifestos for all candidates to National Diet

Example 1: Japanese Campaign Manifestos (Catalinac 2011)

Typical Manifesto:



新宿・千代田・港区の皆さま！ より たしかな 未来を。！ あなたの1票を 生かします！

大つかゆうじ

前衆院土地問題特別委員長
自民党都連政調会長

身近な問題から着実に解決し、

子どもたちの笑顔に満ちた

たしかな日本の未来づくりを目指します。

民主化の進む国際社会において日本の役割はこれまでにない重要な段階を迎えています。いまこそ、たしかな日本の未来の実現のために、グローバルな視点から日本の政治・経済を見直し、国際社会での確たるスタンスを保っていかねばなりません。

そのためには、国内政局の安定を基本に、自らの政治姿勢を正し、'90年代にふさわしい新しい政治を構築せねばなりません。

大つかゆうじは、持てる力の限りを尽くして、たしかな日本の未来の実現のために全力投球の覚悟です。特に、都市議員として、相続税の見直しをはじめ、抜本的な大都市土地政策を確立し、若者に夢と希望を与える住宅をつくることに全力を挙げる所存です。若者の活力こそが、21世紀の日本のよりたしかな未来の実現につながると確信しています。ぜひ、大つかゆうじをご支援ください。

国民の納得する本当の税制改革を！

相続税見直しで定住人口の確保を！

土地有効利用で都心に若者の住宅を！

子どもたちに夢と希望を与える教育の実現！

21世紀を担う子どもたちに夢と希望を／個性と自主性を尊重したきめ細かい教育施策を実施。家庭教育にもスポット。

内需拡大を柱に商店、中小企業の活性化

交通網の整備、都市再開発を推進、多極多心型の都市づくり。先端技術の導入や市場開拓などに緊急融資の実施／

女性新時代の実現

家族と共にある新しい母親像を提言／育児休業、再教育、保育施設の完備など女性が安心して社会参加できる環境を／

「生きがいある第二の人生」づくり

人生80年時代。予防医学、健康増進、成人学校など積極的な人生設計／年金、医療費の安定。寝たきり老人への施策。

全力で取り組んでいます！

私のこれまでの実績など

○地下鉄6・7・12号線建設に大きな貢献／13号線実現に奮闘中／
○都市再開発プロジェクト アークヒルズ建設をはじめ各地に計画を推進／
○市場整備計画の推進と予算化でバックアップ／
○マンション議連会長として永住できる快適な住環境づくりを推進中／
○都心の駐車場対策に尽力、事故撲滅運動を展開交通渋滞解消に全力／

私たちのくらしに、なくてはならない政治家です。

都政、区政と国政のパイプ役としての大つかさんは、抜群の行動派です。鋭い洞察力と明快な判断力を持つ大つかさんは、一区ではもっとも信頼と実績に裏付けられた政治家です。これまで、衆院土地問題に関する特別委員長として、国政レベルの土地施策を積極的に推進する傍ら、若者たちを都心に呼び戻すための、あるいは、高齢社会への各種福祉施策など鋭意邁進中です。

どうぞ、皆さま方のあたたかいご支援を大つかゆうじにお願いいたします。

大つかゆうじさんを推薦します。

鈴木俊一（東京都知事）
山本克忠（新宿区長）
木村 茂（千代田区長）
山田敬治（港区長）

大つかゆうじ 略歴

議員歴 東京都議会議員2期、衆議院議員5期

役職歴 国土政務次官、文部政務次官、衆議院法務委員長、検察官適格審査会委員、自民党調査局長、文教部会長、教育改革に関する調査会会長代理

現役職 前衆議院土地問題特別委員会委員長、党文教制度調査会副会長、党都連政調会長、慶応義塾大学評議員、財国際連合大学協会理事他

Example 1: Japanese Campaign Manifestos (Catalinac 2011)

Japanese Elections:

- Election Administration Commission runs elections $\rightarrow$ district level
- Required to submit manifestos for all candidates to National Diet
- Collected from 1950- 2009

Example 1: Japanese Campaign Manifestos (Catalinac 2011)

Japanese Elections:

- Election Administration Commission runs elections $\rightarrow$ district level
- Required to submit manifestos for all candidates to National Diet
- Collected from 1950- 2009
  - Available only at district level

Example 1: Japanese Campaign Manifestos (Catalinac 2011)

Japanese Elections:

- Election Administration Commission runs elections $\rightarrow$ district level
- Required to submit manifestos for all candidates to National Diet
- Collected from 1950- 2009
  - Available only at district level
  - Until: 2009 national library made texts available on microfilm

Example 1: Japanese Campaign Manifestos (Catalinac 2011)

Japanese Elections:

- Election Administration Commission runs elections $\rightarrow$ district level
- Required to submit manifestos for all candidates to National Diet
- Collected from 1950- 2009
  - Available only at district level
  - Until: 2009 national library made texts available on microfilm
- Collected from microfilm, hand transcribed (no OCR worked), used a variety of techniques to create a TDM

Example 1: Japanese Campaign Manifestos (Catalinac 2011)

Japanese Elections:

- Election Administration Commission runs elections $\rightarrow$ district level
- Required to submit manifestos for all candidates to National Diet
- Collected from 1950- 2009
  - Available only at district level
  - Until: 2009 national library made texts available on microfilm
- Collected from microfilm, hand transcribed (no OCR worked), used a variety of techniques to create a TDM
- Harder for Japanese

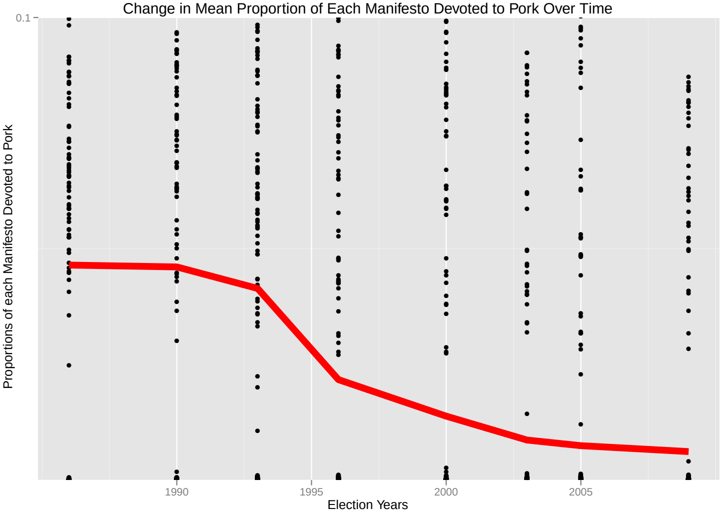
Example 1: Japanese Campaign Manifestos (Catalinac 2014)

- Applies Vanilla LDA (using R Code I'll detail in a moment)
- Output: topics (with Japanese characters)

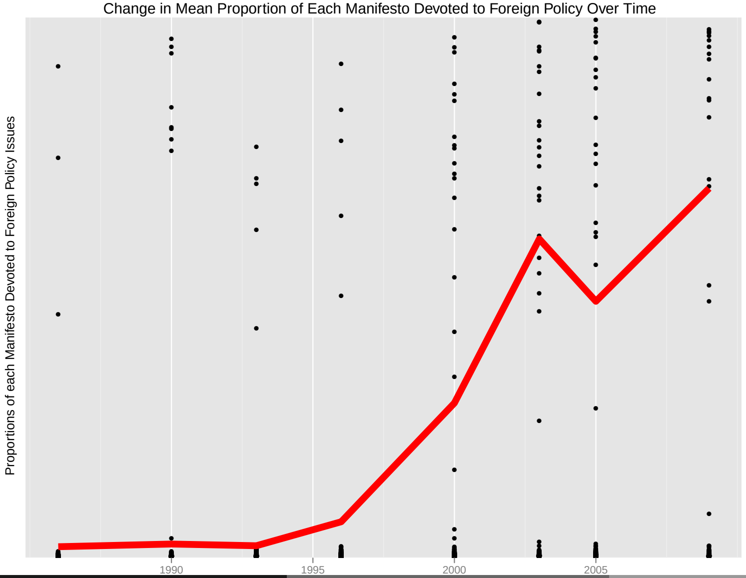
Example 1: Japanese Campaign Manifestos (Catalinac 2011)

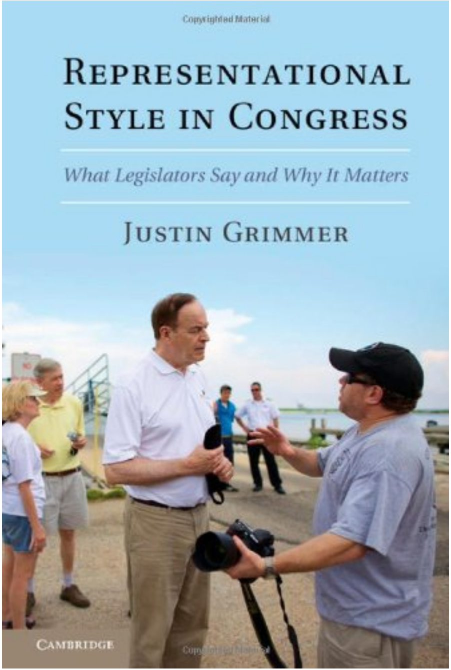
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
改革	年金	推進	区	政治	日本
郵政	円	整備	政策	改革	国
民営	廃止	図る	地域	国民	外交
小泉	改革	つとめる	まち	企業	国家
構造	兆	社会	鹿児島	自民党	社会
政府	実現	対策	全力	日本	国民
官	無駄	振興	選挙	共産党	保障
推進	日本	充実	国政	献金	安全
民	増税	促進	作り	金権	地域
自民党	削減	安定	横浜	党	拉致
日本	一元化	確立	対策	選挙	経済
制度	政権	企業	中小	禁止	守る
民間	子供	実現	発電	憲法	問題
年金	地域	中小	推進	腐敗	北朝鮮
実現	ひと	育成	エネルギー	団体	教育
進める	サラリーマン	制度	企業	区	責任
断行	制度	政治	声	ソ連	力
地方	議員	地域	実現	守る	創る
止める	金	福祉	活性	平和	安心
保障	民主党	事業	自民党	円	目指す
財政	年間	改革	地方	反対	誇り
作る	一掃	確保	尽くす	真	憲法
賛成	郵政	強化	商店	是正	可能
社会	道路	教育	いかす	一掃	道
国民	交代	施設	全国	悪政	未来
公務員	社会保険庁	生活	政党	抜本	ひと
力	月額	支援	ひと	定数	再生
経済	手当	環境	支援	政党	将来
国	談合	発展	経済	金丸	解決
安心	支援	施策	福祉	改悪	基本
Postal privatization	Reducing Wasteful Public Spending	Pork for the District	Policies for the district	Political Reform	Nation

Example 1: Japanese Campaign Manifestos (Catalinac 2011)



Example 1: Japanese Campaign Manifestos (Catalinac 2011)




Copyrighted Material
REPRESENTATIONAL STYLE IN CONGRESS
What Legislators Say and Why It Matters
JUSTIN GRIMMER
CAMBRIDGE

Example 2: Automated Literature Reviews

Recall: literature reviews are hard to conduct

LDA: developed (in part) to help structure `JSTOR` database

Use `JSTOR`'s research service to obtain data to analyze

Question: How do scholars use classic text: Home Style

Analysis: all articles that cite Home Style in JSTOR's data

Example 2: Automated Literature Reviews

Output: topic estimates

- Obtain $\log \theta_k$ from model
- One method to summarize a topic:
    - $\exp(\log \theta_k)$ (select 10-20 biggest words)
    - $\exp(\log \theta_k) - \text{Average}_{j \neq k} \exp(\log \theta_j)$ (select 10-20 biggest words)

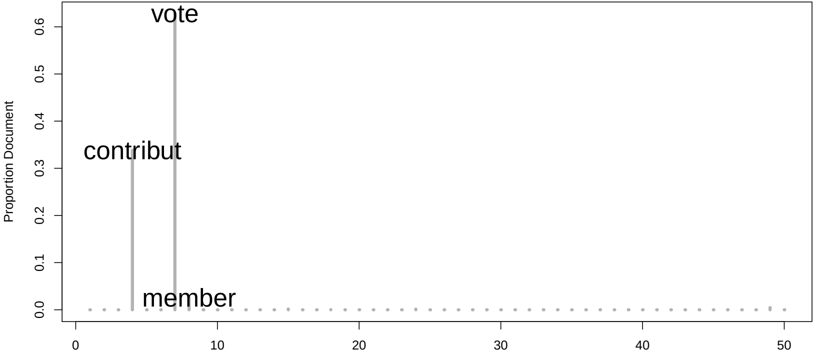
Example 2: Automated Literature Reviews

Example topics:

Label	Stems	Proportic
Life Style	member,district,attent,congress,time,cohort,retir	0.03
Comp.Home	constitu,mp,member,parti,role,local,british	0.02
Casework	casework,district,constitu,variabl,staff,congression,fiorina	0.03
Votes	vote,variabl,model,estim,measur,legisl,constitu	0.04
Id. Shirk	ideolog,vote,shirk,constitu,parti,senat,voter	0.03
C. letters	mail,govern, activ,respond,commun,offic	0.02

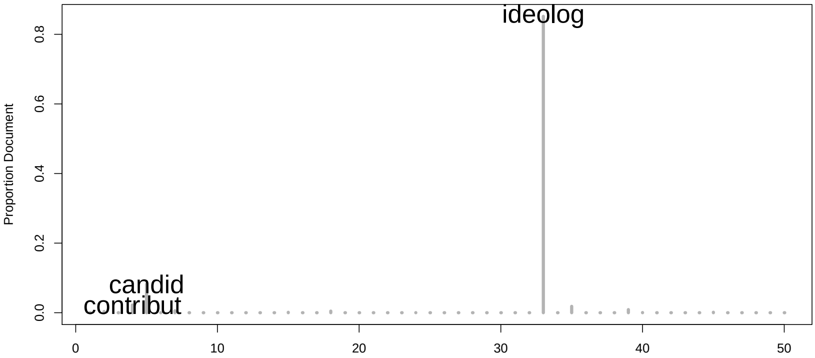
Example Document

Wawro (2001) “A Panel Probit Analysis of Campaign Contributions and Roll Call Votes”



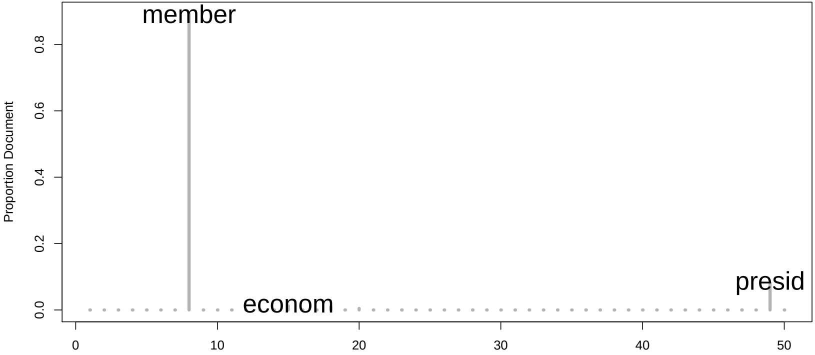
Example Document

Bender (1996) "Legislator Voting and Shirking A Critical Review of the Literature"



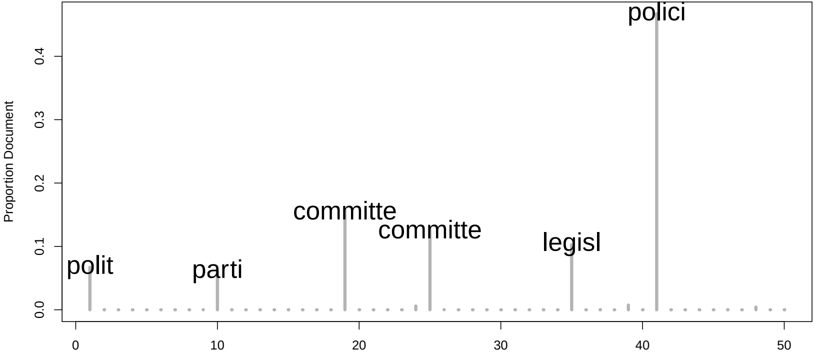
Example Document

Parker (1980) "Cycles in Congressional District Attention"

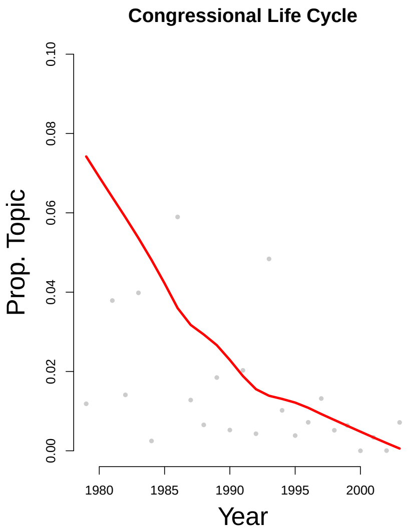


Example Document

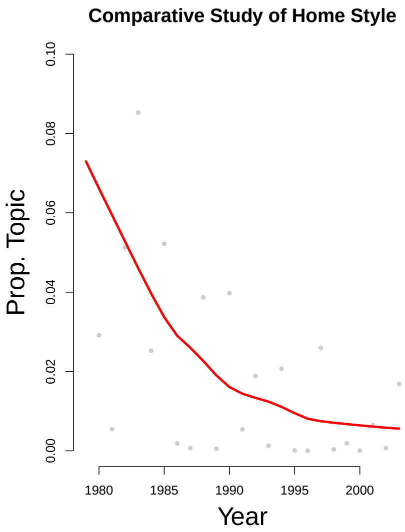
Shepsle (1985) "Policy Consequences of Government by Congressional Subcommittees"



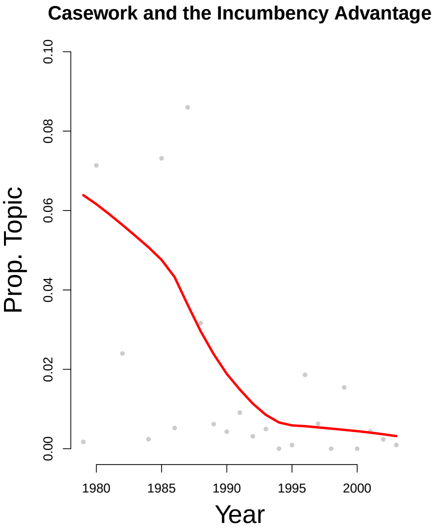
History of Home Style



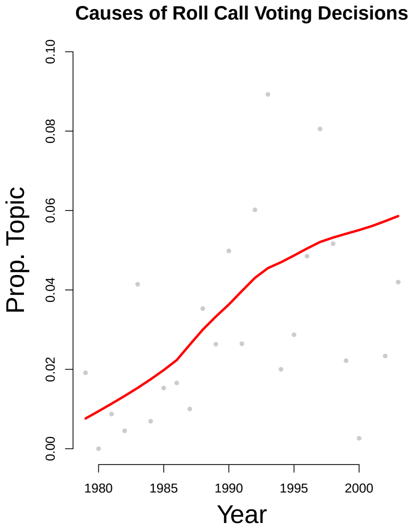
History of Home Style



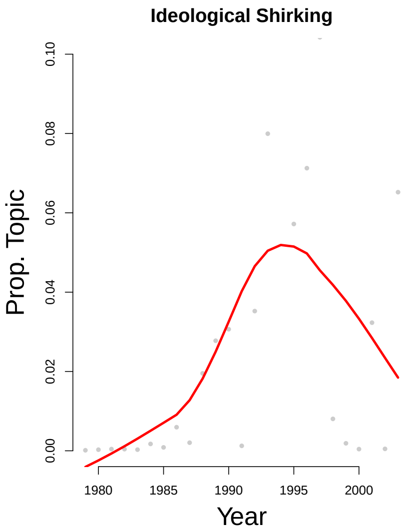
History of Home Style



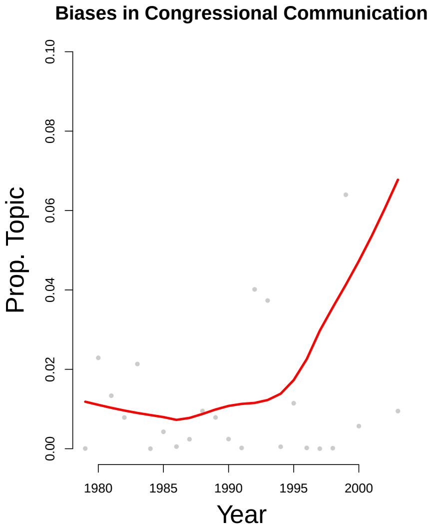
History of Home Style

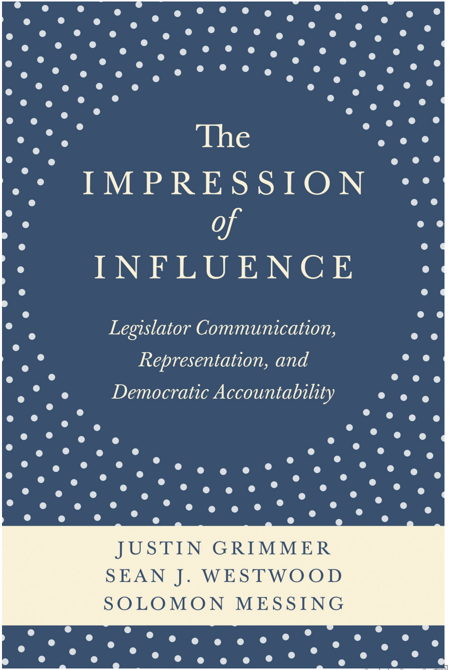


History of Home Style



History of Home Style




The
IMPRESSION
of
INFLUENCE
Legislator Communication,
Representation, and
Democratic Accountability
JUSTIN GRIMMER
SEAN J. WESTWOOD
SOLOMON MESSING

What legislators claim (Grimmer, Westwood, Messing 2014)

What legislators claim (Grimmer, Westwood, Messing 2014) $\rightsquigarrow$ LDA credit claiming press releases

Labels	Key Words	Proportion

What legislators claim (Grimmer, Westwood, Messing 2014) $\rightsquigarrow$ LDA credit claiming press releases

Labels	Key Words	Proportion
Requested appropriations	bill,funding,house,million,appropriations	0.08

"Dave Camp announced today that he was able to secure $2.5 million for widening M-72 from US-31 easterly 7.2 miles to Old M-72. The bill will now head to the Senate for consideration...We have two more hurdles to clear to make sure the money is in the bill when it hits the President's desk: a vote in the Senate and a conference committee" (Camp, 2005)

What legislators claim (Grimmer, Westwood, Messing 2014) $\rightsquigarrow$ LDA credit claiming press releases

Labels	Key Words	Proportion
Requested appropriations	bill,funding,house,million,appropriations	0.08

"Congressman Doc Hastings has boosted federal funding for work on the Columbia Basin water supply for next year. Hastings has added $400,000 for work on the Odessa Subaquifer, which when combined with the funding in the President's budget request, totals $1 million for Fiscal Year 2009"…"Hastings' funding for the Odessa Subaquifer and Potholes Reservoir was included in the Fiscal Year 2009 Energy and Water Appropriations bill which was approved today by the full House Appropriations Committee. (Hastings, 2008)"

What legislators claim (Grimmer, Westwood, Messing 2014) $\rightsquigarrow$ LDA credit claiming press releases

Labels	Key Words	Proportion
Requested appropriations	bill,funding,house,million,appropriations	0.08
Fire department grants	fire,grant,department,program,firefighters	0.08

"Maurice Hinchey (D-NY) today announced that the West Endicott Fire Company has been awarded a $17,051 federal grant to purchase approximately 10 sets of protective clothing, as well as radio equipment and air packs for its volunteer firefighters" (Hinchey, 2008)

What legislators claim (Grimmer, Westwood, Messing 2014) $\rightsquigarrow$ LDA credit claiming press releases

Labels	Key Words	Proportion
Requested appropriations	bill,funding,house,million,appropriations	0.08
Fire department grants	fire,grant,department,program,firefighters	0.08

"Congressman Pete Visclosky today announced that the Crown Point Fire Department will receive a $16,550 Department of Homeland Security (DHS) grant to purchase a modular portable video system" (Visclosky, 2008)

What legislators claim (Grimmer, Westwood, Messing 2014) $\rightsquigarrow$ LDA credit claiming press releases

Labels	Key Words	Proportion
Requested appropriations	bill,funding,house,million,appropriations	0.08
Fire department grants	fire,grant,department,program,firefighters	0.08
Stimulus	recovery,funding,jobs,information, act,	0.06

What legislators claim (Grimmer, Westwood, Messing 2014) $\rightsquigarrow$ LDA credit claiming press releases

Labels	Key Words	Proportion
Requested appropriations	bill,funding,house,million,appropriations	0.08
Fire department grants	fire,grant,department,program,firefighters	0.08
Stimulus	recovery,funding,jobs,information, act,	0.06
Transportation	transportation,project,airport,transit,million	0.06

Correlated Topic Models

Dirichlet distribution$\rightsquigarrow$ Assumes negative covariance between topics
Logistic Normal Distribution$\rightsquigarrow$ Allows some positive covariance between topics

$$
\begin{aligned}
\boldsymbol{\theta}_k &\sim \text{Dirichlet}(\mathbf{1}) \\
\boldsymbol{\eta}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma} &\sim \text{Multivariate Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
\boldsymbol{\pi}_i &= \frac{\exp(\boldsymbol{\eta}_i)}{\sum_{k=1}^{K} \exp(\eta_{ik})} \\
\boldsymbol{\tau}_{im}|\boldsymbol{\pi}_i &\sim \text{Multinomial}(1, \boldsymbol{\pi}_i) \\
x_{im}|\boldsymbol{\theta}_k, \tau_{imk} = 1 &\sim \text{Multinomial}(1, \boldsymbol{\theta}_k)
\end{aligned}
$$

Vanilla Topic Models

1) Vanilla Topic Models
2) Structural Topic Models $\rightsquigarrow$ Different paths for validations

Appendix: Estimating LDA

1) Variational Approximation
2) Collapsed Gibbs Sampling

Variational Approximation

Basic set up

Variational Approximation

Basic set up

Call $q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})$ an arbitrary distribution the approximating distribution.

Variational Approximation

Basic set up
Call $q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})$ an arbitrary distribution the approximating distribution.
Recall $p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}|\boldsymbol{X})$ is the posterior

Variational Approximation

Basic set up

Call $q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})$ an arbitrary distribution the approximating distribution.

Recall $p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}|\boldsymbol{X})$ is the posterior

Our goal is to make $q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T})$ as close as possible to $p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \alpha|\boldsymbol{X})$.

Variational Approximation

Basic set up
Call $q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})$ an arbitrary distribution the approximating distribution.
Recall $p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}|\boldsymbol{X})$ is the posterior
Our goal is to make $q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T})$ as close as possible to $p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \alpha|\boldsymbol{X})$.

$$q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})^* = \arg\min_{q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})} \text{KL}(q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}))||p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}|\boldsymbol{X})))$$

Variational Approximation

Basic set up
Call $q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})$ an arbitrary distribution the approximating distribution.
Recall $p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}|\boldsymbol{X})$ is the posterior
Our goal is to make $q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T})$ as close as possible to $p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}|\boldsymbol{X})$.

$$q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})^* = \arg\min_{q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})} \mathrm{KL}(q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}))||p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}|\boldsymbol{X})))$$

KL is the Kullback-Leibler Divergence between $q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})$ and $p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}|\boldsymbol{X})$.

Variational Approximation

Basic set up

Call $q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})$ an arbitrary distribution the approximating distribution.

Recall $p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}|\boldsymbol{X})$ is the posterior

Our goal is to make $q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T})$ as close as possible to $p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}|\boldsymbol{X})$.

$$q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})^{*} = \arg\min_{q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})} \text{KL}(q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}))||p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}|\boldsymbol{X})))$$

KL is the Kullback-Leibler Divergence between $q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})$ and $p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}|\boldsymbol{X})$.

$$\text{KL}(q||p) = -\sum_{\boldsymbol{T}} \iint q(\boldsymbol{T}, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\alpha}) \log \left\{ \frac{p(\boldsymbol{T}, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\alpha}|\boldsymbol{X})}{q(\boldsymbol{T}, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\alpha})} \right\} d\boldsymbol{\pi} d\boldsymbol{\theta}$$

Variational Approximation

Basic set up

Call $q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})$ an arbitrary distribution the approximating distribution.

Recall $p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}|\boldsymbol{X})$ is the posterior

Our goal is to make $q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T})$ as close as possible to $p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}|\boldsymbol{X})$.

$$q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})^* = \arg\min_{q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})} \mathrm{KL}(q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}))||p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}|\boldsymbol{X})))$$

KL is the Kullback-Leibler Divergence between $q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})$ and $p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}|\boldsymbol{X})$.

$$\mathrm{KL}(q||p) = -\sum_{\boldsymbol{T}} \int\int q(\boldsymbol{T}, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\alpha}) \log \left\{ \frac{p(\boldsymbol{T}, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\alpha}|\boldsymbol{X})}{q(\boldsymbol{T}, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\alpha})} \right\} d\boldsymbol{\pi} d\boldsymbol{\theta}$$

KL-divergence measures dissimilarity between two distributions.

Variational Approximation

Variational Approximation

$$q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})^{*} \quad = \quad \text{arg min}_{q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})} \text{KL}(q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}) || p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha} | \boldsymbol{X}))$$

Variational Approximation

Variational Approximation

$$q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})^* = \text{arg min}_{q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})} \text{KL}(q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}) || p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha} | \boldsymbol{X}))$$

No assumptions about q

Variational Approximation

Variational Approximation

$$q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})^* \quad = \quad \text{arg min}_{q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})} \text{KL}(q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}) || p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha} | \boldsymbol{X}))$$

No assumptions about q then, $q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})^* = p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha} | \boldsymbol{X})$

Variational Approximation

Variational Approximation

$$q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})^* = \arg\min_{q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})} \text{KL}(q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}) || p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha} | \boldsymbol{X}))$$

No assumptions about q then, $q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})^* = p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha} | \boldsymbol{X})$
Simplifying Assumption: $q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}) \equiv q(\boldsymbol{\pi}) q(\boldsymbol{\theta}) q(\boldsymbol{T}) q(\boldsymbol{\alpha})$.

Variational Approximation

Variational Approximation

$$q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})^{*} = \arg\min_{q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})} \text{KL}(q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}) || p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha} | \boldsymbol{X}))$$

No assumptions about q then, $q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})^{*} = p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha} | \boldsymbol{X})$
Simplifying Assumption: $q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}) \equiv q(\boldsymbol{\pi}) q(\boldsymbol{\theta}) q(\boldsymbol{T}) q(\boldsymbol{\alpha})$.
Sufficient to make inference tractable

Variational Approximation

Variational Approximation

$$q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})^* = \arg\min_{q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})} \text{KL}(q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}) || p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha} | \boldsymbol{X}))$$

No assumptions about q then, $q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})^* = p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha} | \boldsymbol{X})$
Simplifying Assumption: $q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}) \equiv q(\boldsymbol{\pi})q(\boldsymbol{\theta})q(\boldsymbol{T})q(\boldsymbol{\alpha})$.
Sufficient to make inference tractable!

Variational Approximation

Variational Approximation

$$q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})^* \quad = \quad \text{arg min}_{q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})} \text{KL}(q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}) || p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha} | \boldsymbol{X}))$$

No assumptions about q then, $q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})^* = p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha} | \boldsymbol{X})$
Simplifying Assumption: $q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}) \equiv q(\boldsymbol{\pi}) q(\boldsymbol{\theta}) q(\boldsymbol{T}) q(\boldsymbol{\alpha})$.
Sufficient to make inference tractable!
So, how do we minimize KL-divergence with respect to $q(\boldsymbol{\pi}) q(\boldsymbol{\theta}) q(\boldsymbol{T}) q(\boldsymbol{\alpha})$?

Variational Approximation

Variational Approximation

$$q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})^{*} = \text{arg min}_{q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})} \text{KL}(q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}) || p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha} | \boldsymbol{X}))$$

No assumptions about q then, $q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})^{*} = p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha} | \boldsymbol{X})$

Simplifying Assumption: $q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}) \equiv q(\boldsymbol{\pi}) q(\boldsymbol{\theta}) q(\boldsymbol{T}) q(\boldsymbol{\alpha})$.

Sufficient to make inference tractable!

So, how do we minimize KL-divergence with respect to $q(\boldsymbol{\pi}) q(\boldsymbol{\theta}) q(\boldsymbol{T}) q(\boldsymbol{\alpha})$?

We solve an equivalent maximization problem

$\log p(\boldsymbol{Y})$

$$\log p(\boldsymbol{Y}) \quad = \quad \log \sum_{\boldsymbol{T}} \iint p(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\alpha}) d\boldsymbol{\theta} d\boldsymbol{\pi}$$

$$
\begin{aligned}
\log p(\boldsymbol{Y}) &= \log \sum_{\boldsymbol{T}} \iint p(\boldsymbol{X}, \boldsymbol{T}, \pi, \theta, \alpha) d\boldsymbol{\theta} d\boldsymbol{\pi} \\
&= \log \sum_{\boldsymbol{T}} \iint \frac{q(\pi, \theta, \boldsymbol{T}, \alpha)}{q(\pi, \theta, \boldsymbol{T}, \alpha)} p(\boldsymbol{X}, \boldsymbol{T}, \pi, \theta) d\boldsymbol{\theta} d\boldsymbol{\pi}
\end{aligned}
$$

$$
\begin{aligned}
\log p(\boldsymbol{Y}) &= \log \sum_{\boldsymbol{T}} \iint p(\boldsymbol{X}, \boldsymbol{T}, \pi, \theta, \alpha) d\boldsymbol{\theta} d\boldsymbol{\pi} \\
&= \log \sum_{\boldsymbol{T}} \iint \frac{q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \alpha)}{q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \alpha)} p(\boldsymbol{X}, \boldsymbol{T}, \pi, \theta) d\boldsymbol{\theta} d\boldsymbol{\pi} \\
&\geq \sum_{\boldsymbol{T}} \iint q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \alpha) \log \left\{ \frac{p(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\pi}, \boldsymbol{\theta}, \alpha)}{q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \alpha)} \right\} d\boldsymbol{\pi} d\boldsymbol{\theta}
\end{aligned}
$$

$$
\begin{aligned}
\log p(\boldsymbol{Y}) &= \log \sum_{\boldsymbol{T}} \iint p(\boldsymbol{X}, \boldsymbol{T}, \pi, \theta, \alpha) d\boldsymbol{\theta} d\boldsymbol{\pi} \\
&= \log \sum_{\boldsymbol{T}} \iint \frac{q(\pi, \theta, \boldsymbol{T}, \alpha)}{q(\pi, \theta, \boldsymbol{T}, \alpha)} p(\boldsymbol{X}, \boldsymbol{T}, \pi, \theta) d\boldsymbol{\theta} d\boldsymbol{\pi} \\
&\geq \sum_{\boldsymbol{T}} \iint q(\pi, \theta, \boldsymbol{T}, \alpha) \log \left\{ \frac{p(\boldsymbol{X}, \boldsymbol{T}, \pi, \theta, \alpha)}{q(\pi, \theta, \boldsymbol{T}, \alpha)} \right\} d\boldsymbol{\pi} d\boldsymbol{\theta}
\end{aligned}
$$

$$
\sum_{\boldsymbol{T}} \iint q(\pi, \theta, \boldsymbol{T}, \alpha) \log \left\{ \frac{p(\boldsymbol{X}, \boldsymbol{T}, \pi, \theta, \alpha)}{q(\pi, \theta, \boldsymbol{T}, \alpha)} \right\} d\boldsymbol{\pi} d\boldsymbol{\theta} = \mathcal{L}(q)
$$

$$\log p(\boldsymbol{Y}) =$$

$$\log p(\boldsymbol{Y}) = \mathrm{KL}(q||p) + \mathcal{L}(q)$$

$$\underbrace{\log p(\boldsymbol{Y})}_{\text{Fixed Number}} \quad = \quad \mathrm{KL}(q||p) + \mathcal{L}(q)$$

$$\underbrace{\log p(\boldsymbol{Y})}_{\text{Fixed Number}} = \mathrm{KL}(q||p) + \mathcal{L}(q)$$

If $\mathcal{L}(q)$ increases

$$\underbrace{\log p(\boldsymbol{Y})}_{\text{Fixed Number}} = \text{KL}(q||p) + \mathcal{L}(q)$$

If $\mathcal{L}(q)$ increases then $\text{KL}(q||p)$ must decrease.

$$\underbrace{\log p(\boldsymbol{Y})}_{\text{Fixed Number}} = \text{KL}(q||p) + \mathcal{L}(q)$$

If $\mathcal{L}(q)$ increases then $\text{KL}(q||p)$ must decrease.
Choose q to maximize $\mathcal{L}(q)$

$$\underbrace{\log p(\boldsymbol{Y})}_{\text{Fixed Number}} \quad = \quad \text{KL}(q||p) + \mathcal{L}(q)$$

If $\mathcal{L}(q)$ increases then $\text{KL}(q||p)$ must decrease.
Choose q to maximize $\mathcal{L}(q)$ equivalent to minimizing $\text{KL}(q||p)$.

Iterative algorithm to maximize $\mathcal{L}(q)$.

Iterative algorithm to maximize $\mathcal{L}(q)$.
Initialize $q(\boldsymbol{\pi})^{\text{old}}$, $q(\boldsymbol{\theta})^{\text{old}}$, $q(\boldsymbol{T})^{\text{old}}$, and $q(\boldsymbol{\alpha}^{\text{old}})$

Iterative algorithm to maximize $\mathcal{L}(q)$.
Initialize $q(\boldsymbol{\pi})^{\text{old}}$, $q(\boldsymbol{\theta})^{\text{old}}$, $q(\boldsymbol{T})^{\text{old}}$, and $q(\boldsymbol{\alpha}^{\text{old}})$
Find $q(\boldsymbol{\pi})^{\text{new}}$ to max $\mathcal{L}(q)$– holding $q(\boldsymbol{\theta})^{\text{old}}$, and $q(\boldsymbol{T})^{\text{old}}$ constant

Iterative algorithm to maximize $\mathcal{L}(q)$.
Initialize $q(\boldsymbol{\pi})^{\text{old}}$, $q(\boldsymbol{\theta})^{\text{old}}$, $q(\boldsymbol{T})^{\text{old}}$, and $q(\boldsymbol{\alpha}^{\text{old}})$
Find $q(\boldsymbol{\pi})^{\text{new}}$ to max $\mathcal{L}(q)$– holding $q(\boldsymbol{\theta})^{\text{old}}$, and $q(\boldsymbol{T})^{\text{old}}$ constant
Find $q(\boldsymbol{\theta})^{\text{new}}$ to max $\mathcal{L}(q)$– holding $q(\boldsymbol{T})^{\text{old}}$, and $q(\boldsymbol{\pi})^{\text{new}}$ constant

Iterative algorithm to maximize $\mathcal{L}(q)$.
Initialize $q(\boldsymbol{\pi})^{\text{old}}$, $q(\boldsymbol{\theta})^{\text{old}}$, $q(\boldsymbol{T})^{\text{old}}$, and $q(\boldsymbol{\alpha}^{\text{old}})$
Find $q(\boldsymbol{\pi})^{\text{new}}$ to max $\mathcal{L}(q)$– holding $q(\boldsymbol{\theta})^{\text{old}}$, and $q(\boldsymbol{T})^{\text{old}}$ constant
Find $q(\boldsymbol{\theta})^{\text{new}}$ to max $\mathcal{L}(q)$– holding $q(\boldsymbol{T})^{\text{old}}$, and $q(\boldsymbol{\pi})^{\text{new}}$ constant
Find $q(\boldsymbol{T})^{\text{new}}$ to max $\mathcal{L}(q)$– holding $q(\boldsymbol{\theta})^{\text{new}}$, and $q(\boldsymbol{\pi})^{\text{new}}$ constant

Iterative algorithm to maximize $\mathcal{L}(q)$.
Initialize $q(\boldsymbol{\pi})^{\text{old}}$, $q(\boldsymbol{\theta})^{\text{old}}$, $q(\boldsymbol{T})^{\text{old}}$, and $q(\boldsymbol{\alpha}^{\text{old}})$
Find $q(\boldsymbol{\pi})^{\text{new}}$ to max $\mathcal{L}(q)$– holding $q(\boldsymbol{\theta})^{\text{old}}$, and $q(\boldsymbol{T})^{\text{old}}$ constant
Find $q(\boldsymbol{\theta})^{\text{new}}$ to max $\mathcal{L}(q)$– holding $q(\boldsymbol{T})^{\text{old}}$, and $q(\boldsymbol{\pi})^{\text{new}}$ constant
Find $q(\boldsymbol{T})^{\text{new}}$ to max $\mathcal{L}(q)$– holding $q(\boldsymbol{\theta})^{\text{new}}$, and $q(\boldsymbol{\pi})^{\text{new}}$ constant
Assume $q(\boldsymbol{\alpha})$ is degenerate, maximization step

Iterative algorithm to maximize $\mathcal{L}(q)$.
Initialize $q(\boldsymbol{\pi})^{\text{old}}$, $q(\boldsymbol{\theta})^{\text{old}}$, $q(\boldsymbol{T})^{\text{old}}$, and $q(\boldsymbol{\alpha}^{\text{old}})$
Find $q(\boldsymbol{\pi})^{\text{new}}$ to max $\mathcal{L}(q)$– holding $q(\boldsymbol{\theta})^{\text{old}}$, and $q(\boldsymbol{T})^{\text{old}}$ constant
Find $q(\boldsymbol{\theta})^{\text{new}}$ to max $\mathcal{L}(q)$– holding $q(\boldsymbol{T})^{\text{old}}$, and $q(\boldsymbol{\pi})^{\text{new}}$ constant
Find $q(\boldsymbol{T})^{\text{new}}$ to max $\mathcal{L}(q)$– holding $q(\boldsymbol{\theta})^{\text{new}}$, and $q(\boldsymbol{\pi})^{\text{new}}$ constant
Assume $q(\boldsymbol{\alpha})$ is degenerate, maximization step
Guaranteed convergence: $\mathcal{L}(q)$ is convex in $q(\boldsymbol{\pi})$, $q(\boldsymbol{\theta})$,and $q(\boldsymbol{T})$

Finding $q(\boldsymbol{\pi})^{\text{new}}$.

Finding $q(\boldsymbol{\pi})^{\text{new}}$.

$$\mathcal{L}(q) = \int q(\boldsymbol{\pi})^{\text{new}} \underbrace{\left\{ \sum_{\boldsymbol{T}} \int \log p(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\theta}, \boldsymbol{\pi}) q(\boldsymbol{T}, \boldsymbol{\alpha})^{\text{old}} q(\boldsymbol{\theta})^{\text{old}} d\boldsymbol{\theta} \right\}}_{\text{E}_{\boldsymbol{T},\boldsymbol{\theta}}[\log p(\boldsymbol{Y},\boldsymbol{T},\boldsymbol{\pi},\boldsymbol{\theta})]} d\boldsymbol{\pi}$$

$$- \int q(\boldsymbol{\pi})^{\text{new}} \log q(\boldsymbol{\pi})^{\text{new}} d\boldsymbol{\pi} + \text{constants}$$

Finding $q(\boldsymbol{\pi})^{\text{new}}$.

$$\mathcal{L}(q) = \int q(\boldsymbol{\pi})^{\text{new}} \underbrace{\left\{ \sum_{\boldsymbol{T}} \int \log p(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\theta}, \boldsymbol{\pi}) q(\boldsymbol{T}, \boldsymbol{\alpha})^{\text{old}} q(\boldsymbol{\theta})^{\text{old}} d\boldsymbol{\theta} \right\}}_{\mathrm{E}_{\boldsymbol{T},\boldsymbol{\theta}}[\log p(\boldsymbol{Y},\boldsymbol{T},\boldsymbol{\pi},\boldsymbol{\theta})]} d\boldsymbol{\pi}$$

$$- \int q(\boldsymbol{\pi})^{\text{new}} \log q(\boldsymbol{\pi})^{\text{new}} d\boldsymbol{\pi} + \text{constants}$$

$$\log \tilde{p}(\boldsymbol{\pi}) = \mathrm{E}_{\boldsymbol{T},\boldsymbol{\theta}}[\log p(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\alpha})] + \text{constants}$$

Substituting in $\log \tilde{p}(\boldsymbol{\pi})$,

$$= \int q(\boldsymbol{\pi})^{\text{new}} \log \left(\frac{\tilde{p}(\boldsymbol{\pi})}{q(\boldsymbol{\pi})^{\text{new}}} \right) d\boldsymbol{\pi}$$

Substituting in $\log \tilde{p}(\boldsymbol{\pi})$,

$$
\begin{aligned}
&= \int q(\boldsymbol{\pi})^{\text{new}} \log\left(\frac{\tilde{p}(\boldsymbol{\pi})}{q(\boldsymbol{\pi})^{\text{new}}}\right) d\boldsymbol{\pi} \\
&= -\text{KL}(q(\boldsymbol{\pi})^{\text{new}} || \tilde{p}(\boldsymbol{\pi}))
\end{aligned}
$$

Substituting in $\log \tilde{p}(\boldsymbol{\pi})$,

$$\begin{aligned} &= \int q(\boldsymbol{\pi})^{\text{new}} \log \left(\frac{\tilde{p}(\boldsymbol{\pi})}{q(\boldsymbol{\pi})^{\text{new}}} \right) d\boldsymbol{\pi} \\ &= -\text{KL}(q(\boldsymbol{\pi})^{\text{new}} || \tilde{p}(\boldsymbol{\pi})) \end{aligned}$$

At a maximum when $q(\boldsymbol{\pi})^{\text{new}} = \tilde{p}(\boldsymbol{\pi})$

Substituting in $\log \tilde{p}(\boldsymbol{\pi})$,

$$
\begin{aligned}
&= \int q(\boldsymbol{\pi})^{\text{new}} \log\left(\frac{\tilde{p}(\boldsymbol{\pi})}{q(\boldsymbol{\pi})^{\text{new}}}\right) d\boldsymbol{\pi} \\
&= -\text{KL}(q(\boldsymbol{\pi})^{\text{new}} || \tilde{p}(\boldsymbol{\pi}))
\end{aligned}
$$

At a maximum when $q(\boldsymbol{\pi})^{\text{new}} = \tilde{p}(\boldsymbol{\pi})$
Equivalently,

$$
\begin{aligned}
\log q(\boldsymbol{\pi})^{\text{new}} &= \log \tilde{p}(\boldsymbol{\pi}) \\
&= \text{E}_{\boldsymbol{T},\boldsymbol{\theta}}[\log p(\boldsymbol{X}, \boldsymbol{T}, \pi, \boldsymbol{\theta}, \boldsymbol{\alpha})] + \text{constants}
\end{aligned}
$$

Substituting in $\log \tilde{p}(\boldsymbol{\pi})$,

$$
\begin{aligned}
&= \int q(\boldsymbol{\pi})^{\text{new}} \log\left(\frac{\tilde{p}(\boldsymbol{\pi})}{q(\boldsymbol{\pi})^{\text{new}}}\right) d\boldsymbol{\pi} \\
&= -\text{KL}(q(\boldsymbol{\pi})^{\text{new}} || \tilde{p}(\boldsymbol{\pi}))
\end{aligned}
$$

At a maximum when $q(\boldsymbol{\pi})^{\text{new}} = \tilde{p}(\boldsymbol{\pi})$
Equivalently,

$$
\begin{aligned}
\log q(\boldsymbol{\pi})^{\text{new}} &= \log \tilde{p}(\boldsymbol{\pi}) \\
&= \text{E}_{\boldsymbol{T},\boldsymbol{\theta}}[\log p(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\alpha})] + \text{constants}
\end{aligned}
$$

Or,

Substituting in $\log \tilde{p}(\boldsymbol{\pi})$,

$$
\begin{aligned}
&= \int q(\boldsymbol{\pi})^{\text{new}} \log \left(\frac{\tilde{p}(\boldsymbol{\pi})}{q(\boldsymbol{\pi})^{\text{new}}} \right) d\boldsymbol{\pi} \\
&= -\text{KL}(q(\boldsymbol{\pi})^{\text{new}} || \tilde{p}(\boldsymbol{\pi}))
\end{aligned}
$$

At a maximum when $q(\boldsymbol{\pi})^{\text{new}} = \tilde{p}(\boldsymbol{\pi})$
Equivalently,

$$
\begin{aligned}
\log q(\boldsymbol{\pi})^{\text{new}} &= \log \tilde{p}(\boldsymbol{\pi}) \\
&= \text{E}_{\boldsymbol{T},\boldsymbol{\theta}}[\log p(\boldsymbol{X}, \boldsymbol{T}, \pi, \boldsymbol{\theta}, \boldsymbol{\alpha})] + \text{constants}
\end{aligned}
$$

Or,

$$
q(\boldsymbol{\pi})^{\text{new}} = \frac{\exp\left\{\text{E}_{\boldsymbol{T},\boldsymbol{\theta}}[\log p(\boldsymbol{X}, \boldsymbol{T}, \pi, \boldsymbol{\theta}, \boldsymbol{\alpha})]\right\}}{\int \exp\left\{\text{E}_{\boldsymbol{T},\boldsymbol{\theta}}[\log p(\boldsymbol{X}, \boldsymbol{T}, \pi, \boldsymbol{\theta}, \boldsymbol{\alpha})]\right\} d\boldsymbol{\pi}}
$$

Algorithm

Algorithm
Initialize $q(\boldsymbol{\pi})^{\text{old}}$, $q(\boldsymbol{\theta})^{\text{old}}$, and $q(\boldsymbol{T})^{\text{old}}$

Algorithm
Initialize $q(\boldsymbol{\pi})^{\text{old}}$, $q(\boldsymbol{\theta})^{\text{old}}$, and $q(\boldsymbol{T})^{\text{old}}$

$$\log q(\boldsymbol{\pi})^{\text{new}} = \mathrm{E}_{\boldsymbol{T},\boldsymbol{\theta}}[\log p(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\alpha})] + \text{constants}$$

Algorithm
Initialize $q(\boldsymbol{\pi})^{\text{old}}$, $q(\boldsymbol{\theta})^{\text{old}}$, and $q(\boldsymbol{T})^{\text{old}}$

$$\begin{aligned}
\log q(\boldsymbol{\pi})^{\text{new}} &= \mathrm{E}_{\boldsymbol{T},\boldsymbol{\theta}}[\log p(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\alpha})] + \text{constants} \\
\log q(\boldsymbol{\theta})^{\text{new}} &= \mathrm{E}_{\boldsymbol{T},\boldsymbol{\pi}}[\log p(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\alpha})] + \text{constants}
\end{aligned}$$

Algorithm
Initialize $q(\boldsymbol{\pi})^{\text{old}}$, $q(\boldsymbol{\theta})^{\text{old}}$, and $q(\boldsymbol{T})^{\text{old}}$

$$\begin{aligned}
\log q(\boldsymbol{\pi})^{\text{new}} &= \text{E}_{\boldsymbol{T},\boldsymbol{\theta}}[\log p(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\alpha})] + \text{constants} \\
\log q(\boldsymbol{\theta})^{\text{new}} &= \text{E}_{\boldsymbol{T},\boldsymbol{\pi}}[\log p(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\alpha})] + \text{constants} \\
\log q(\boldsymbol{T})^{\text{new}} &= \text{E}_{\boldsymbol{\theta},\boldsymbol{\pi}}[\log p(\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\alpha})] + \text{constants}
\end{aligned}$$

Algorithm

Initialize $q(\boldsymbol{\pi})^{\text{old}}$, $q(\boldsymbol{\theta})^{\text{old}}$, and $q(\boldsymbol{T})^{\text{old}}$

$$\begin{aligned}
\log q(\boldsymbol{\pi})^{\text{new}} &= \mathrm{E}_{\boldsymbol{T},\boldsymbol{\theta}}[\log p(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\alpha})] + \text{constants} \\
\log q(\boldsymbol{\theta})^{\text{new}} &= \mathrm{E}_{\boldsymbol{T},\boldsymbol{\pi}}[\log p(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\alpha})] + \text{constants} \\
\log q(\boldsymbol{T})^{\text{new}} &= \mathrm{E}_{\boldsymbol{\theta},\boldsymbol{\pi}}[\log p(\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\alpha})] + \text{constants}
\end{aligned}$$

All expectations over approximating distribution.

Algorithm

Initialize $q(\boldsymbol{\pi})^{\text{old}}$, $q(\boldsymbol{\theta})^{\text{old}}$, and $q(\boldsymbol{T})^{\text{old}}$

$$
\begin{aligned}
\log q(\boldsymbol{\pi})^{\text{new}} &= \text{E}_{\boldsymbol{T},\boldsymbol{\theta}}[\log p(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\alpha})] + \text{constants} \\
\log q(\boldsymbol{\theta})^{\text{new}} &= \text{E}_{\boldsymbol{T},\boldsymbol{\pi}}[\log p(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\alpha})] + \text{constants} \\
\log q(\boldsymbol{T})^{\text{new}} &= \text{E}_{\boldsymbol{\theta},\boldsymbol{\pi}}[\log p(\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\alpha})] + \text{constants}
\end{aligned}
$$

All expectations over approximating distribution.

To compute, rely on factorization in posterior

Variational Approximation Update Steps

Carrying out the expectations, we obtain the following forms $\rightsquigarrow$ derivation not assumption

$$
\begin{aligned}
q(\boldsymbol{\tau}_{im}) &= \text{Multinomial}(1, \boldsymbol{r}_{im}) \\
q(\boldsymbol{\theta}_k) &= \text{Dirichlet}(\boldsymbol{\eta}_k) \\
q(\boldsymbol{\pi}_i) &= \text{Dirichlet}(\gamma_i)
\end{aligned}
$$

Update for $q(\boldsymbol{\tau}_{im})$

Consider document i, word m, topic k

$$
\begin{aligned}
r_{imk} &\propto \exp\left(I(x_{im} = j)E[\log \eta_{kj}] + E[\log \gamma_{ik}]\right) \\
&\propto \exp\left(I(x_{im} = j)\left[\Psi(\eta_{kj}) - \Psi(\sum_{l=1}^{J} \eta_{kl})\right] + \Psi(\gamma_{ik}) - \Psi(\sum_{m=1}^{K} \gamma_{im}))\right)
\end{aligned}
$$

where $\Psi(\cdot)$ is the digamma function (the derivative of the gamma function)

Update for $q(\boldsymbol{\eta}_k)$

Consider word j and topic k, then,

$$\eta_{jk} \quad \propto \quad 1 + \sum_{i=1}^{N} \sum_{m=1}^{M_i} r_{imk} x_{im}$$

Update for $q(\gamma_i)$

Consider word topic k and document i

$$\gamma_{ik} \quad \propto \quad \alpha_k + \sum_{m=1}^{M_i} r_{imk}$$

Update for α)

Fast Newton-Raphson Algorithm